Speech@Home: An Exploratory Study

A.J. Brush
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
ajbrush@microsoft.com

Paul Johns
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
pauljoh@microsoft.com

Kori Inkpen
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
kori@microsoft.com

Brian Meyers
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
brianme@microsoft.com

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

.

Abstract
To understand how people might use a speech dialog system in the public areas of their homes, we conducted an exploratory field study in six households. For two weeks each household used a system that logged motion and usage data, recorded speech diary entries and used Experience Sampling Methodology (ESM) to prompt participants for additional examples of speech commands. The results demonstrated our participants' interest in speech interaction at home, in particular for web browsing, calendaring and email tasks, although there are still many technical challenges that need to be overcome. More generally, our study suggests the value of using speech to enable a wide range of interactions.

Keywords
Speech, home technology, diary study, domestic

ACM Classification Keywords
H.5.2 User Interfaces: Voice I/O.

General Terms
Human Factors, Experimentation

Introduction
The potential for using speech in home environments has long been recognized. Smart home research has suggested using speech dialog systems for controlling

home infrastructure (e.g. [5, 8, 10, 12]), and a variety of novel home applications use speech interfaces, such as cooking support systems [2] and the Nursebot personal robotic assistant for the elderly [14]. Typically these projects have identified a particular problem or domain, such as home automation, and used speech as one possible input method.

In contrast, the goal of this study was to understand more generally how people might use speech input to interact with computers in the public spaces of their homes for all the different types of tasks they do throughout their day. Past research has shown that some households place computers in public spaces including kitchens and living rooms and that these computers are often shared [3]. We refer to shared computers located in public spaces in the home as home kiosks. We wanted to understand how people interact with these devices and whether or not speech interaction would be effective and desirable. We hypothesized that speech might allow users to interact with the kiosk from a distance or while multi-tasking. For example, while busy cooking one could call out "Add milk to the shopping list" from across the room.

While speech recognition technology (e.g., Dragon Naturally Speaking software [6] or Windows Speech Recognition [18]) has vastly improved in recent years, it is still most successful when trained for use by a single person using a close-talk microphone, or in cases where the set of potential commands is limited (e.g., hands-free dialing of cell phones). Home environments, with their background noise, multiple people, and range of possible tasks, remain a challenging environment for spoken dialog systems.

We conducted a two week exploratory speech diary study in six households to understand whether or not speech interaction was desirable, and, if so, what types of tasks and styles of interaction would participants be interested in. We built a speech diary computer (SDC) using a Dell All-in-One PC running Windows 7. The SDC not only provided general purpose computing resources, but allowed participants to record speech diary entries by saying the phrase "*Speech Command*" followed by their desired command, i.e., "*Speech Command, open email*", or "*Speech Command, what is traffic like?*" When our software detected participants saying the command phrase, it saved the next 5 seconds of audio. Although these commands were only recorded and not executed, gathering diary entries in this way provided the opportunity to learn from participants in-situ as they went about their daily lives.

In addition to collecting spontaneous speech diary entries, we also had five specific topics we believed participants might want to use speech to request information on: Weather, Calendar, Email, Traffic and News. During the study we prompted participants at various times to record an entry based on one of these topics. This ESM-based data [1] was used to compare the language used by different people about the same set of topics.

Our results highlight the potential for speech interaction with computers in public spaces in the home; however, many challenges still remain that must be solved before robust speech interaction at home becomes a reality. Our findings will help inform others working on speech interfaces for homes and provide directions for future home kiosk developments.

Related Work

Speech recognition is supported on many operating systems including Windows [18] and Macintosh [9] and applications like Dragon NaturallySpeaking [6]. Often focused on providing accessibility, these systems allow people to control features of their computer and dictate text to the computer. Feng et al., [7] conducted a six month field study of ten participants, half with physical impairments, using speech technology. The results suggested that a number of participants, including all with physical impairments, preferred to use speech for navigation, particularly for web browsing and email. Some participants also used speech for games and programming tasks. While this study provides valuable insights on how people with a wide range of physical abilities adopt and use speech technology, our study explored what participants might want to do without constraining them to what is currently possible.

In home environments, considerable research has explored speech dialog systems, primarily focused on home automation (e.g., [5, 10, 12]). Homes with multiple people and background noise present a number of challenges for these systems, including both technical and interaction design challenges. Automatic speech recognition without close-talk microphones is one of the main technical challenges and many researchers (e.g., [5, 21]) are developing microphone arrays to improve speech recognition in homes.

Nass and Brave [11] report on experiments with speech interfaces that explore the effect of different types of synthetic voices on users' behavior and satisfaction. Their studies consistently demonstrate that people respond to synthetic voices as if the technology is a social actor and interact with it similarly to how they would a human. Wolters et al., [19] analyzed communication between older and younger users of a simulated speech dialog system and identified two groups: a "social" group and a "factual" group that were not related to age. The "social group" was more likely to treat the system as human, for example, using politeness markers.

Although there is a rich history of research on speech dialog systems for homes, we are unaware of any similar studies that seek to understand in-situ what users might wish to do using speech in their home. Given the challenges of working in homes, most studies of smart homes are carried out in laboratory environments (e.g. [8, 10]) with pre-specified tasks. In contrast, our study offers the opportunity to learn from participants in their homes.

Speech Diary Computer

To build the speech diary computer we used Dell Studio One 19 desktop all-in-one computers (see Figure 1) running Windows 7. These computers have an 18.5 inch multi-touch screen, wireless keyboard and mouse, built-in camera, microphone, and speakers. We added a Phidgets 1111 PIR Motion Sensor mounted on top of the computer near the camera (visible in Figure 1C) to avoid being triggered by pets. This sensor detects changes in infrared radiation and can detect motion in a 60 degree field of view within 5 meters. The motion sensor turned off the computer screen after 60 seconds of inactivity. We also installed the Personal Vibe (PV) Windows Activity Logger [13]. PV allows us to track what applications participants used on the computer and the length of their usage.

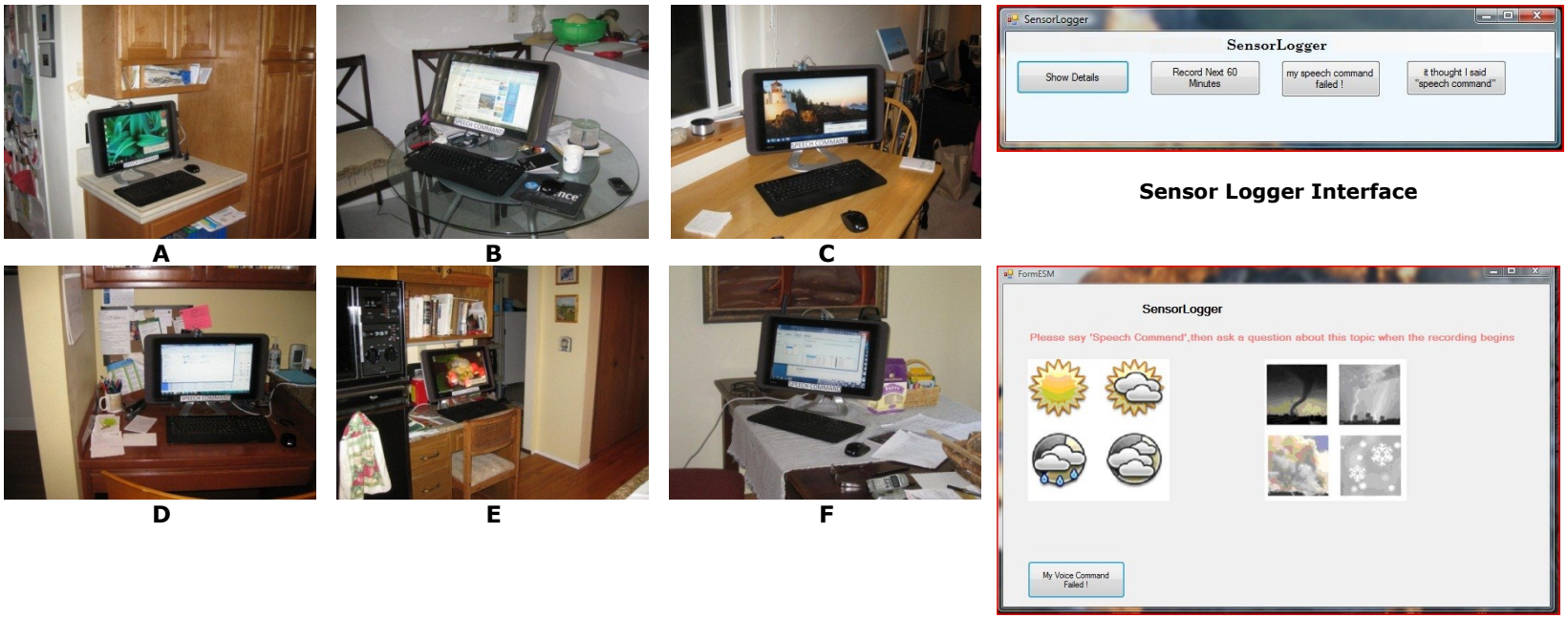

Figure 1. Speech Diary Computers deployed in each of the six households (A-F), the Sensor Logger interface, and an example ESM dialog about weather.

Our Sensor Logger application continuously listened to audio in the environment and collected speech diary entries and motion data. We used the Microsoft Speech API and customized the grammar so that it only attempted to recognize the phrase "*Speech Command*". We chose a multiple word prompt to improve recognition as single word commands like "computer" are too common in everyday speech. "*Speech Command*" had the best recognition rate among an initial test group. When the application recognized this phrase, it chimed and then recorded the next 5 seconds of audio as a diary entry before giving a second chime.

The Sensor Logger interface, shown in Figure 1, had four buttons. The "my speech command failed!" button was used to indicate when the application had trouble recognizing the "Speech Command" phrase. The "it thought I said speech command" button was used to mark false positives such as when the application chimed to indicate recognition, but the participant did not say "Speech Command." The "Record Next 60 Minutes" button was used to gather additional background audio data. The "Show Details" button was used to show logging history and allow playback and/or deletion of previously recorded audio (no participants took advantage of this option).

We included ESM-style prompting to collect entries about specific topics across participants and to increase the number of speech entries recorded by participants. Although the ESM prompts could potentially bias the spontaneous entries, given the exploratory nature of the study, we felt that gathering data using multiple methods would give us richer insights.

Sensor Logger displayed ESM dialogs (Figure 1) at least two hours apart and only after observing 20 seconds of motion. Prompts stopped for the day after four responses had been recorded or twelve dialogs had been shown. Each ESM dialog showed one of six different topics: Weather, Calendar, Email, News, Traffic, or a question mark image to indicate participants could ask about anything they wanted. We used pictures to describe the topics so that we would not bias the participants toward particular words.

As previous studies have highlighted, there are considerable privacy concerns when recording audio and video data in people's homes [e.g., 15] so we took care to minimize what we recorded. Although Sensor Logger was always listening, it only saved the recorded audio in three circumstances. First, when the application recognized the "Speech Command" phrase, a five second audio snippet and 30 second video snippet (with audio) were recorded. The video was recorded to help disambiguate false positives. Second, when a participant pushed the "my speech command failed" button, 60 seconds of audio before and after the event was saved, and lastly when ESM dialogs were presented to the user, audio was saved while the ESM window was open and a reminder chime was sounded each minute. The window closed after a diary entry was made, a participant closed the window manually, or eight minutes had passed without a diary entry. We also let participants know the location of the videos in case they wanted to review and delete any video files.

Study Method

We conducted the study in an urban region in the Northwest United States during the summer of 2009. Six households with a total of 22 members participated. The median age of the adults was 44; the median age of the children was 14. Participants' occupations varied and included homemaker, music teacher, finance manager, and software document manager. All households were recruited to have three or more people, own two or more computers, and have no members that regularly use a speech interface (such as Dragon Naturally Speaking) so that they would not be biased by current state of the art in speech interfaces.

Due to our primary interest in the potential for using speech to interact with home kiosks, we recruited households that currently had a computer in a public space, which is used by two or more people, at least once a week. Each household used the speech diary computer for two weeks. We visited each household 3 times during the study. During our first visit, participants filled out a background form about the computers in their home, how they are used, and their locations. We interviewed each member of the household about their past experience using speech interfaces and then gave each participant 10 minutes to write down ways they might want to use speech to interact with a computer in their home. We encouraged participants to record any ideas they had, even if the idea seemed impossible using existing technology.

We found participants had typically used Interactive Voice Response systems (IVR) such as calling customer support for a bank. Speech interaction using a mobile phone was less common, although 11 participants had some experience. Five participants had experience using speech with a computer. Four of the five reported brief experiments with either Windows Speech Recognition or Dragon Naturally Speaking, while one had used speech recognition more extensively four years ago. Three of these participants reported being satisfied, one was neutral, and one was very dissatisfied. Other experiences with speech interfaces included six participants, mostly kids, who had used speech to interact with video games.

During the first visit, we also installed the three speech diary computers we had built in households A, B, C. We configured the systems to use the household's internet connection and then demonstrated how to record a diary entry by saying "Speech Command." We explained the Sensor Logger interface and had each household member try the interaction until their command was recognized. We showed participants the detail window where they could play back the audio files we were recording, see the motion logging information, and push the record button to record additional background audio data. We also explained the ESM popup dialogs. Finally, we showed participants the video that was recorded (30 seconds around each speech entry) so they would be aware of the cameras' field of view. We followed up with each household by email four days after our visit to check-in and make sure the study was proceeding smoothly. Due to ongoing interest in how households use computers, we also installed PV logger on two other computers in the household when technically feasible. We did not use this additional PV data in this study.

At our second visits, we collected the speech diary computers from A, B, C and installed them in D, E, F. Participants in A, B, C filled out a post-survey about their experience, which we reviewed with each participant in a semi-structured interview, asking additional questions and clarifying feedback as needed. In our third visit we collected the systems and post-study feedback from D, E, F and PV logging data from the additional computers from all households

For their participation each household received four software gratuities (max value $600 each). Participants could receive up to two additional gratuities. One for allowing us to configure PV to record the URLs of websites they visited and a second for recording at least 10 hours of additional background audio data (using the "Record Next 60 minutes" button in Figure 2). The additional audio data was gathered to provide realistic background noise levels for future work and was not analyzed for this study. All households chose to provide the additional data and received six gratuities.

Results

Many participants were enthusiastic about the potential for speech interfaces. On the post-survey, 12 participants reported that if the applications they felt were most important worked with speech they would be "Very Interested" in owning a computer with a speech interface; while 7 more were "Somewhat Interested." The remaining three participants were less interested in using speech to interact with the computer, highlighting speech recognition failures, and

their perception that the efficiency of keyboard interaction makes speech unnecessary.

Assessing Opportunities for Using Home Kiosks
The first part of our analysis focused on how much each family used the SDC and the amount of opportunity for speech interaction. Figure 1 shows the placement of each SDC. Households A, D, and E placed the computers on desk space in their kitchens, while households B, C, F, placed the computers on tables in the kitchen which were not typically used for meals.

We evaluated the motion data, the PV activity data and the post interview questionnaire. Unfortunately a technical problem prevented us from gathering some of the PV data and thus we do not have activity data for families A and F. Table 1 shows the average number of hours each family used the SDC and the amount of motion detected around the computer.

House	A	B	C	D	E	F	Avg.
Avg. Hrs. of PC Activity per Day	*	6.0	2.0	2.9	0.2	*	2.8
Avg. Hrs. of Motion per Day	4.5	7.5	3.4	7.4	2.8	1.8	4.6^{+}
Avg. Hrs. of Motion With No PC Activity	*	1.5	1.4	4.5	2.6	*	2.5

Table 1. Motion and computer logging data.

Families A, C and D reported that they often used the SDC instead of another computer in the house. The logging data for families C and D showed between two and three hours of usage per day. Family B used the SDC considerably more, with an average of 6 hours of usage per day. In contrast, families E and F reported that they rarely used the SDC. Family E had only 10 minutes of activity per day as the computer was inadvertently put on the wrong wireless network (a neighbors') which made it too slow to be usable. This family did not have sufficient need or desire for the computer to notify us or to change it themselves. In Household F, F_F58 (participant ids denote household, gender and age) reported that she rarely used the SDC but she tried to force herself to use it for the study

The motion sensor data (Table 1) showed that the families did spend time near the computer, with the amount of activity ranging from an average of almost 2 hours per day (Family F) to more than 7 hours per day (Family B and D). Although we do not expect that household members should be interacting with technology all of the time they are in their kitchen or living room, this data suggests that speech diary computers were placed into spaces where household members are spending considerable amounts of time and that during much of this time participants would be capable of seeing the screen or speaking a command.

To assess if participants might be interested in using speech interaction at a distance, we looked at the data for speech entries when no motion was detected (i.e., the participant was not in view of the computer and likely not in range of other input options). We found 14 instances where speech entries were made when there was no motion registered for 15 seconds before the entry which demonstrates some interest in using speech interaction at a distance from the kiosk.

We examined session length information from PV to understand how long participants used the SDC at a time. By establishing a session break whenever the

computer was inactive for 5 minutes we found that across the families there were 771 sessions. The minimum session length was 1 second. Examining the activity in these extremely short sessions indicated the user was often waking the screen after it had blanked. The longest session was slightly over 4 hours, with an average length of 13.3 minutes and a median of 5 minutes. The high percentage of short sessions suggests that participants may be using the computer to find information snacks – small, easy to digest, chunks of information that can be easily attended to in the course of other activities.

Tasks for Speech Interaction

One of the goals of the study was to understand in-situ which tasks, if any, participants would want to perform using speech to interact with a home kiosk.

Pre-study Interview

At the beginning of the study we asked each participant to write down what types of speech interactions they would like to have their home computer support. In total, participants reported 164 different ideas. We examined participants' suggestions and grouped them into four high-level categories. Using these categories, two researchers independently coded all of the speech ideas and then discussed any differences in order to come to agreement. Table 2 shows the categorization of participants' initial ideas and the sub-categories.

1. Web: Ideas that involved using a web browser to go to a specific site or request a search for information. Sub-categories include Navigation, Search, News, Directions, and Weather. These categories were inspired by previous work by [4, 17]. The Navigational sub-category with web browser commands was the most common (e.g. "D_M48: *goto chase.com*", "E_M18: *web browsing*", "A_M47: *bring up website*"). Participants also suggested search requests such as A_F44 "*what time stores open*," and "A_M5: *is my Jonas Brothers CD in yet at the library*."

2. Communication/PIM: Ideas related to communication and personal information management (PIM). Sub-categories include Email, synchronous communication through IM or phone calls, lists, and calendar, which includes reminders. The Calendar/Reminder sub-category was most popular (11% of all ideas). Example ideas in this category include: "C_F25: *enter event on a personal calendar*", "D_F15: *open calendar*", and "B_F30: *remind me about my appointments*").

3. Other Applications: Ideas related to using particular applications (excluding communications, PIM, and web browser). Sub-categories included Games, Multimedia (e.g. F_M64:"*play song*"), and Documents (C_F25: "*highlight paragraphs*"). In the Applications category, ideas related to multimedia (e.g. "D_F12: *go to youtube*") were most popular, but interacting with other applications including documents and photos were suggested by multiple people.

4. Control: Ideas related to controlling general computer behavior (Computer) or home automation systems (Home). Examples include turning on or off the computer and turning on a house alarm. If participants were specific about opening or working with a particular application covered by another sub-category (e.g. "D_F12: *go to email*"), we classified the idea in that category (e.g., Email). However, if the idea was expressed generally (e.g., "C_M26: *open programs and software*"), it was classified in the Control-Computer

Table 2. Categorization of participants' initial speech ideas and their speech diary entries. *The other and social interation categories were not present in the speech ideas and were added based on new ideas that emerged from the diary entries. 18 entries were placed in two categories.

Categories		Initial Ideas	Speech Entries
Web	Navigational	11 (7%)	45 (11%)
	Search	11 (7%)	37 (9%)
	Weather	3 (2%)	20 (5%)
	Directions/Traffic	3 (2%)	11 (3%)
	News/RSS	2 (1%)	4 (1%)
	Total	**30 (18%)**	**117 (30%)**
Communications/ PIM	Calendar/Reminder	18 (11%)	42 (11%)
	Email	9 (5%)	29 (7%)
	Lists	12 (7%)	16 (4%)
	Call/IM	5 (3%)	6 (2%)
	Contacts	4 (2%)	3 (1%)
	Total	**48 (29%)**	**96 (24%)**
Control	Computer	37 (23%)	26 (7%)
	Home	2 (1%)	34 (9%)
	Help	*	10 (3%)
	Total	**39 (24%)**	**70 (18%)**
Applications	Multimedia	13 (8%)	15 (4%)
	Documents	9 (5%)	5 (1%)
	Photos	7 (4%)	1 (0%)
	Games	6 (4%)	14 (4%)
	Calculator	4 (2%)	2 (1%)
	Recipes	5 (3%)	13 (3%)
	Other Apps	3 (2%)	4 (1%)
	Total	**47 (29%)**	**54 (14%)**
Other		*	11 (3%)
Social Interaction		*	48 (12%)
Total		164	396 (378 unique)

sub-category. Participants' ideas in the Control category focused on using speech to open and close programs (11), to request help (4), start up and shut down the computer (3), and log-on (3).

Speech Diary Entries

In analyzing the audio data we categorized a total of 378 five second audio clips as speech diary entries. This included 49 entries recorded in response to the "anything" category of the ESM dialog. Again, two researchers independently coded the entries according to the coding scheme developed to analyze the pre-study interview. Three additional categories were added: Help, Other, and Social Interaction. There are a total of 396 categorized items, because 18 entries were put in two categories, see Table 2.

Analyzing the responses reinforced to us the value of conducting the study in-situ. Compared to the ideas suggested by participants in the pre-study, the field deployment collected more realistic input as participants typically made more specific requests. For example, in the Web-Search sub-category, entries included "A_M47: *Search for the latest smart phones from Verizon*" and "D_F44: *check soccer schedule through LWISA*" compared to pre-study ideas: "D_F14: *say words to search on Bing*" or "F_M21: *define words*." Specificity also contributed to the dramatic decline in the percentage of responses classified in the Control-Computer sub-category to 7% compared to 23% of pre-study ideas. Rather than generic computer commands (e.g. "C_M26: *open programs and software*"), participants were more specific in diary entries. For example, "C_M23: *open up windows media player*." In addition, during the diary study a participant might repeat the same entry at different times unlike

the unique list of ideas participants provided in the pre-study interviews. This gives us a more realistic picture of the frequency that participants might issue certain types of requests.

Overall, the Web (30%) and Communications/PIM (24%) categories dominated in participants' spontaneous diary entries. The Web-Navigational sub-category had 11% of the diary entries overall, showing interest in using speech to open, close and issue specific navigational commands to a web browser application (e.g. "F_F58: *return to job.com*"). More general search requests, the Web sub-category, accounted for 9% of diary entries. The Calendar/Reminder sub-category (11% overall) and Email related requests (7%) were the most commonly made requests in the Communications/PIM category.

In the Control category (18%), the popularity of the Control-Home sub-category in the diary entries (9% increase compared to pre-study) was due entirely to members of Household A who were interested in managing their house alarm using speech. 65% of their requests (21 of 32) were to turn the house alarm on or off. The percentage of diary entries in the applications category declined relative to pre-study entries (from 29% to 14%). This may have been related to the fact that participants generally did not add their own content (music, documents) to the SDC so they may have been less motivated to make diary entries related to application use.

Help related entries (e.g. "B_M37: *tell me why I can't get connected*") appeared more frequently in the diary data so we added a sub-category to Control. We also added "Other" to handle the few cases (11) where the entries did not directly fit with any other category (e.g. "CM_23: *I'm hungry, go get me some food*").

Speech Task Preferences

Although we gathered both spontaneous diary data and prompted speech topic data, our results suggest that participants' spontaneous responses were likely not biased by the ESM prompts. For example, the two speech topics that participants indicated were least relevant to them, traffic and news, also had very small numbers of speech entries (traffic = 3% overall, news =1%). Conversely, speech topics participants indicated were most relevant to them: Email, Weather and Calendar had higher percentages of entries (Email = 7%, Weather = 5%, Calendar = 11%).

Therefore by comparing participants' speech entries and feedback on the post-survey a fairly clear picture of participants' preferred tasks for using speech emerges. In particular, the speech entries demonstrate interest in web browsing related requests both for explicit navigation (e.g. "C_M23: *open up craigslist.org*") and more general search queries (e.g. "A_F44: *find recipe for chocolate chip cookies*"). Tasks that involved using speech for calendaring and reminding tasks (e.g. "B_M23: *remind me my appointment at 5:00*") as well as email (e.g. "F_F58: *open gmail, open sent mail*"), also occurred frequently in participant's speech entries. Post-survey data also supports participants' interest in calendaring. When asked for the three most important applications or features that a speech interface should support, Calendar was the most common response, stated by 8 participants.

Finally, to compare with participants' current computer usage, we examined the applications recorded by PV.

We found that over the two week period our participants used Web browsers 62% of the time. They spent 17% of the time using Office productivity tools like Word or Excel and 6% of the time using a local Email client like Outlook. This extensive use of the web correlates well with 30% of participants' speech entries being web and search related.

It is important to note that the method we used to gather speech entries certainly influenced the types of entries that participants made and likely encouraged single commands or short phrases since there was a 5 second limit. The five second limit to diary entries was a challenge observed by some participants. As F_F58 said "*5 seconds is sometimes too little time to say all you want to say*" and we noticed that recordings by Household E were frequently cut-off.

While a working system might encourage longer speech interactions, several participants expressed that speech input was slower than using the keyboard and mouse. For example, F_M64 commented mid-way through the study "*This* [speech] *is not an appealing or useful feature*," and felt that for the things he wanted to do, the keyboard was most effective. Thus, we believe that it is important for home kiosks to focus on using speech not to replace keyboard and mouse interaction, but to augment it; e.g., supporting speech when users are at a distance from the computer or in conjunction with other input modalities. For example, allowing the user to say "*open Facebook*" as they approached a computer and then using the keyboard to enter a status update.

We were also interested to note that several of the common pre-study ideas and diary entries recorded focused on opening applications ("F_F58: *open Internet Explorer*"). This is functionality currently available using existing speech technology (e.g. [9, 18]). Changing the user experience to make it easier for users to use speech when it is most appropriate and in combination with other input modalities may help users take advantage of functionality that already exists.

Lastly, smart home research often assumes speech interaction throughout the home. Eight of our participants reported that they thought of things to say to the SDC at times when they were not nearby. Ten participants were very interested in being able to use speech throughout their home rather than just near the computer, and two participants were uninterested.

Social Interaction with the Home Kiosk

Although none of the pre-study ideas included social aspects that one might observe in human interactions (e.g., greetings or politeness markers), every household except F had at least one diary entry related to social interaction. Analysis of the speech entries revealed that 12% (48) included language that indicated the participant was treating the computer as a social actor. For example, several of the participants used politeness markers, ("*please*" or "*thank you*"), included a greeting ("*Good morning*", "*bye*") or used a personal pronoun ("*see you later when I get home*"), language similar to that used by the social group identified by Wolters et al. [19]. Additionally, 32 of these actions only involved social interaction, and did not have another command associated with it.

Households A (24 of 48 entries with social presence) and C (13) had the most entries categorized in social presence. The children in Household A, ages 8 and 5, were responsible for most of their household's diary

entries involving social presence (19 out of 24) and delighted in saying "hi" and "goodbye" to the SDC. Nine of the speech topic responses from three households (A, C, F) also included words of politeness ("*please*", "*thank you*").

Past research by Nass and his colleagues [11] has shown people treat computers as social actors, particularly when the system includes speech output. Similarly, in their deployment of the Tableau system, which created and displayed pictures based on sensor data collected in kitchens, Pousman et al. [15] also had participants name their devices. However, given those system were reacting and interacting with the user, we were somewhat surprised to the extent that our participants treated the SDC as a social actor since it was essentially a recording device that did not act on their commands.

Four households mentioned wanting to replace the phrase "Speech Command" with a name they selected. For example, D_F44 said "*I would prefer to choose a fun name for our computer (like Roscoe)*" and "*when you give it a name, makes it more special, between a dog and a family member.*" Allowing households to personalize the phrase used to initiate speech interaction represents an additional challenge for speech recognition, but was clearly desired by some.

Speech Grammar

We used the speech topic data gathered using ESM to gain insight into how similar participants' requests were about the same topic. 175 speech topic responses were recorded for the following topics: Calendar (36), Weather (31), Email (39), News (40), and Traffic (29). We examined the responses to see how often a key word appeared. Traffic and Email responses were the most consistent. Responses to traffic prompts included the word "traffic" appearing at some point in 90% of the responses, while for responses to the Email topic, 72% had the word "email," and 21% "mail," which covered 93% of the entries. The word "weather" appeared in 68% of the Weather responses and "temperature" was found in 10%. In contrast, responses to the Calendar and News topic were most diverse. The word calendar appeared in only 47% of responses to that category with "day" accounting for 22% more. The word "news" appeared in only 38% of the responses to the News topic and "headline" appeared in 20% more.

We also observed that speech topic responses from people within a household differed (e.g. A_M47: "*please, bring up my email*" and A_F44: "*open email*"). More surprising, we observed different styles of interactions from the same participant and a variety of different phrases for the same question. For example, Participant A_M47 consistently asked for the top five news stories in response to News dialogs, but used five different phrases: "*what are the top 5 news articles today?*", "*show me the top 5 news articles of the day*," "*bring up today's 5 top articles*," and "*bring up the 5 top news stories today*." This data suggests that grammars to support speech interactions may need to be quite flexible or that the system makes it evident what phrases are understood.

SPEECH AND HOME KIOSK CHALLENGES

While many of our participants were enthusiastic about the potential for speech interaction on a home kiosk, our study also made it clear that many challenges remain. When we deployed the SDCs we expected that

the speech dialog system would have some challenges related to recognition of the phrase "Speech Command", particularly given the environmental challenges and that the Microsoft Speech API is optimized for a single person using a close-talk microphone. However, we were disappointed by an extremely high number of false positives, times when the recognizer thought that "Speech Command" was said but it had not been. During the study we collected 3005 5-second audio recordings when the computer thought Speech Command was said, of which 87% were false positives.

This extreme level of false positives did not occur in pilots in our own homes, but we found several homes in our study had higher levels of background noise. Many false positives in B (96% false positives) and E (97%) seemed to be caused by audio from the TV or radio which were on frequently. D (81%) also had a problem with false positives, which we believe was caused by often having multiple people in the kitchen (with 4 teenage girls, it was a busy place). Accents were also problematic. For B, recognition of the "Speech Command" phrase was particularly problematic. While B_M37's English was excellent, it was not his native language and the rest of his family was less comfortable with English. His wife, B_F30 observed on the post-survey that it was "*very hard for computer to get my accent*."

Initiating the speech interaction was also somewhat challenging for participants. Overall, participant satisfaction was mixed concerning use of the phrase "Speech Command" to start interaction (8 were somewhat dissatisfied, 9 were somewhat satisfied, and 1 was very satisfied). Nine participants reported that on average they only needed to say the "Speech Command" one or two times. However, 11 participants reported that they needed to say the phrase three or more times before the computer would recognize it. Reasons why participants were not satisfied were primarily related to recognition, for example, "E_M61: *have to repeat*". However, other reasons included that the phrase was "A_M47: *a bit long*" and a "B_M37: *A bit of a mouthful*." Because the speech recognition engine tries to improve recognition by using a gender-specific speech recognition profile, some participants of a different gender than the person who used the system reported additional problems. For example, on the post-survey C_F25's indicated recognition was worse for her than her two male roommates and that the computer "*almost never recognized my command*."

Although we are excited by our results, it is important to also acknowledge the strengths and limitations of our study method. We gathered data in people's homes as they went about their everyday lives; however this necessarily limited the number of households to six, all of whom where in the same geographic area. Our methodology required household members to imagine interactions they would like to do using speech and not all participants were equally inspired to contribute diary entries, particularly those from households B (6% of entries), E (7%) and F (10%) who had a range of problems from bad recognition, slow internet access, and less excitement for speech interaction.

Discussion

Our results highlight both the potential for speech interaction with computers in public spaces in the home and the challenges. Despite the fact that the system we provided did not actually act on any of the speech input

and had an extremely high number of false positives, on the post-survey many participants still reported a desire for speech as part of a home kiosk interface.

The data participants provided suggests promising initial applications to build for those interested in exploring speech interaction in home environments. Our participants were particularly interested in using speech input for controlling web browsing applications, for requesting information and for calendaring and email tasks. While an understanding of the tasks for which people are interested in using speech is critical to successfully incorporate speech interaction into home kiosks, we are particularly interested in the types of interactions enabled by speech and believe our study suggests two valuable directions for future research: supporting kiosk interactions with speech and exploring the roll of social interaction.

Supporting Kiosk Interactions with Speech

Using speech allows a person to interact with the computer without needing to be near enough to touch it. This enables a range of additional types of interactions including: while one approaches the computer, is at a distance from the computer, or even when the screen is out of sight of the user.

We found support for the utility of these types of interactions from the data we collected. First, several of participants' pre-study and speech diary entries were for short and targeted interactions that initiate a task. These commands could be issued when approaching the computer, thus readying the system for use when the person arrives. Using speech in conjunction with other input modalities to extend the distance at which one can interact with the computer and supporting a seamless transition to traditional keyboard and mouse input may feel quite natural to participants. In addition, using speech to "jump start" an interaction makes speech useful even for people who are very proficient at using the keyboard and mouse. In their interaction framework developed for public displays, Vogel and Balakrishnan [20] highlighted the importance of a smooth transition between types of interaction as a user approaches an ambient display, and similarly we feel that speech interactions with kiosks must allow the user to transition smoothly from speech interaction to other types of interaction as they approach the display.

Second, we have some evidence that participants had opportunity and interest in using speech to interact with the computer at a distance. The motion sensor data highlighted times when participants were within sight of the computer, but not actively using it. In addition, some participants made speech entries when they were not within view of the motion sensor and thus likely not able to view the computer display. While we collected motion sensor data, going forward, home kiosks that sense proximity data would be valuable to determine how far the person is from the computer and support their speech interaction accordingly. For example, if the person is far away, but can still see the computer screen, any visual feedback during the interaction will need to be large enough to be visible. Or if the display is not visible, feedback will need to be audible. These styles of interaction: approach, at a distance, and out of sight, have implications for the type of proximity sensing, microphones, and audio output capabilities that home kiosks will require.

The usage logging data gathered on the SDC suggested people may be grabbing information snacks, small

chunks of information, from the computer. We believe speech could help support information snacking behavior. Being able to say "Speech Command, check my email" while in the process of cooking is much more efficient than pausing cooking, starting a program and logging into your email; many of these requests were for personal information suggesting the value of having speaker identification support in a home kiosk.

From "Speech Command" to "Sara"
While several researchers are already exploring social response to computer interfaces (e.g., [11, 15, 16, 19]), results from our study indicate users desire to treat the home kiosk as a social actor.

Although highly desired, allowing households to select a specific name (e.g., "*Sara*") instead of a known phrase ("*Speech Command*") complicates speech recognition as different combinations of names, speakers and environments likely have different false positive rates. Fortunately, since participants stressed the importance of minimizing false positives, we believe participants would be satisfied choosing from a set of easily recognized names. It is also technically possible for each member of the household to use a different name. In discussion with households, this did not seem to hold much appeal, but might be interesting to explore.

More generally, when considering how a home kiosk should respond and interact with people, the combination of speech and multiple people raises interesting directions for future research. Past research on interactions with synthetic voices (e.g., [11, 16]), has focused on one person interacting with the interface and it is not obvious how recommendations made for one-to-one interaction (e.g., people are more positively oriented to a synthetic voice of the same gender) generalize to households.

Concluding Remarks

The data we gathered in our field study provides insight into how participants want to use speech to interact with a home kiosk. Our findings identify applications that appeal to participants (e.g. Traffic, Email) where participants used consistent language which might be easy to recognize. More generally our findings highlight the potential appeal of using speech to enable additional interactions with home kiosks while approaching them, at a distance, and while the kiosk is not visible.

Moving forward we are developing a speech enabled home kiosk that will allow us to conduct research to address both the technical and interaction challenges of building a speech dialog system for the home. We are motivated by the problem space because it is clear that such a system will not succeed without both research into the technical details of speech comprehension and research into the user interaction models. We hope to leverage the unique qualities of the home environment to scope the technical challenges and thus have a usable system long before the full speech recognition of multiple speakers in a noisy environment problem is solved. For example, determining the identity of the speaker would help improve accuracy and this capability may be straight-forward in a home setting where the total number of possible speakers is quite small. We are also interested in exploring how personalization of the speech initiation phrase affects participants' satisfaction with the speech interaction. While considerable research is necessary before robust home kiosks become a reality, the enthusiasm many

participants (including infrequent diary contributors) had for a working system highlights the potential of speech to enable novel interaction in home environments.

References

[1] Barrett, L.F., Barrett D.J.: An Introduction to Computerized Experience Sampling in Psychology. Social Science Computer Review, 19, 2 (2001),175-185.

[2] Bradbury,J., Shell, J., Knowles, C. Hands on cooking: towards an attentive kitchen, Proc. CHI. 2003, 996-997.

[3] Brush, A.J. and Inkpen, K., Yours, Mine, Ours? Sharing and Use of Technology in Domestic Environments, Proc. Ubicomp 2007, 109-126.

[4] Church, K, Smyth, B., Understanding the intent behind mobile information needs, Proc. IUI 2009, 247—256.

[5] Coelho, G.E., Serralheiro, A.J., Netti, J.P., Microphone Array Front-End Interface for Home Automation, Proc. HSCMA 2008, 184-187.

[6] Dragon NaturallySpeaking. www.nuance.com/naturallyspeaking/

[7] Feng, J., Zhu, S., Hu, R., Sears, A., Speech technology in real world environment: early results from a long term study, Proc. Assets '08. 233-234.

[8] Kleindienst, J., Macek, T., Serédi, L., Šedivý, J. Interaction framework for home environment using speech and vision, Image and Vision Computing, V. 25, Issue 12, 2007, 1836-1847

[9] Mac Speakable Items. www.apple.com/accessibility/macosx/physical.html

[10] Möller, S, Krebber, J., Smeele, P. Evaluating the speech output component of a smart-home system. Speech Communications 48 (2006) 1-27.

[11] Nass, C. and Brave, S. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. The MIT Press. 2005.

[12] Oulasvirta A. Engelbrecht, K. Jameson, A., Möller, S, Communication Failures in the speech-based control of smart home systems. Proceedings of the Third International Conference on Intelligent Environments, 2007, 35-143.

[13] Personal Vibe, research.microsoft.com/en-us/downloads/0ea12e49-8b29-4930-b380-a5a00872d229/default.aspx

[14] Pollack, M., Brown, L., Colbry, D., McCarthy, C. E., Orosz, C., Peintner, B., Ramakrishman, S., and Tsamardinos, I. 2003. Autominder: An intelligent cognitive orthotic system for people with memory impairment. Robotics Auton. Syst. 44, 273--282

[15] Pousman, Z., Romero, M., Smith, A., Mateas M., Living with Tableau Machine: a longitudinal investigation of a curious domestic intelligence. Proc. Ubicomp 2008. 370-379.

[16] Reeves, B. & Nass, C. The Media Equation CSLI Publications, 1996.

[17] Sohn , T., Li, K., Griswold, W., Hollan, J. A diary study of mobile information needs, Proc. CHI 2008.

[18] Windows Speech Recognition. www.microsoft.com/enable/products/windowsvista/speech.aspx

[19] Wolters, M., Georgila, K., Moore, J. MacPherson, S., Being Old Doesn't Mean Acting Old: How Older Users Interact with Spoken Dialog Systems, ACM Trans. Access. Computer, 2, 1 Article 2, (May 2009).

[20] Vogel, D. and Balakrishnan, R. Interactive Public Ambient Displays: Transitioning from Implicit to Explicit, Public to Personal, Interaction with Multiple Users. UIST 2004, 137-146.

[21] Ying J.; Yu L.; Yan L.; Kozintsev, I, Distributed Microphone Arrays for Digital Home and Office, Proc. ICASSP 2006. 1065 – 1068Adobe Acrobat Reader 7. http://www.adobe.com/products/acrobat